Apéndice C

Correlación lineal – Significación de parámetros

Objetivos

- ✓ Cuadrados mínimos-datos con errores
- ✓ Incertidumbre de parámetros de un ajuste
- ✓ Significación estadística
- ✓ Bandas de predicción
- ✓ Datos con error en las dos variables

En este apéndice extendemos la discusión de cuadrados mínimos, iniciada en el Cap. 7, al caso en que los datos tengan errores. Se analiza la significación estadística de los parámetros extraídos de un ajuste y sus incertezas. Se presenta una breve digresión de bandas de incertidumbre en una regresión lineal y el caso de una regresión lineal en que las dos variables tienen errores.

C.1 Regresión lineal – Datos con errores

Consideremos el caso de un conjunto de mediciones (X_i, Y_i), donde el error en el valor de Y_i viene dado por σ_i (ver Figura C.1). El objetivo de esta sección es extender la discusión de cuadrados mínimos iniciada en el Cap. 7 al caso de datos con errores. Analizamos el procedimiento que permite obtener la línea que mejor ajusta los datos experimentales y las incertezas asociadas en la determinación de los parámetros del ajuste.

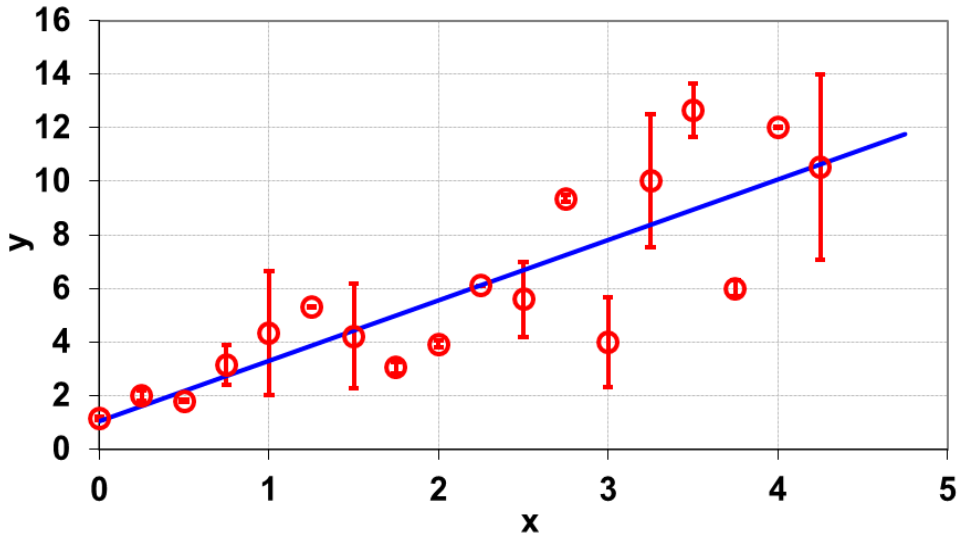

Figura C.1 Representación gráfica de un conjunto de datos experimentales (X_i, Y_i) con errores en el eje Y dados por los valores σ_i. La línea continua azul es la recta obtenida por cuadrados mínimos.

Al igual que lo hicimos en el Cap. 7, suponemos que los datos pueden describirse por la ecuación lineal de la forma:

$$Y(x) = a \cdot x + b \tag{C.1}$$

Definimos el valor de chi cuadrado, χ^2, como:

$$\chi^2 = \sum_{i=1}^{N} w_i \cdot (Y_i - a \cdot x_i - b)^2 \tag{C.2}$$

Aquí w_i es un factor de peso o ponderación que se puede definir de distintos modos según el problema en estudio. Un modo usual de pesar los datos es hacerlo usando sus respectivos errores σ_i del siguiente modo:

$$w_i = \frac{1}{\sigma_i^2} . \tag{C.3}$$

con

$$W = \sum_i w_i \tag{C.4}$$

Si todos los datos tienen igual ponderación, es decir, si w_i=1, entonces W=N = *número total de datos*. Desde luego, la Ec. (C.3) representa sólo una de las tantas formas en que pueden ponderarse los datos. La elección más adecuada de los factores de ponderación depende del problema específico en consideración.

1. El método de cuadrados mínimos consiste en elegir como los mejores valores de a y b aquellos valores que minimicen el valor de χ^2 —Ec. (C.2)—. El resultado de este procedimiento resulta en:[1,2,3,4]

$$a_0 = \frac{W \cdot \left(\sum YX\right) - \left(\sum X\right) \cdot \left(\sum Y\right)}{W \cdot \left(\sum X^2\right) - \cdot \left(\sum X\right)^2} = \frac{< YX > - < X > \cdot < Y >}{\left(< X^2 > - < X >^2\right)} = \frac{Cov(YX)}{S_X^2} , \tag{C.5}$$

y

$$b_0 = \frac{\left(\sum Y\right) \cdot \left(\sum X^2\right) - \left(\sum X\right) \cdot \left(\sum Y \cdot X\right)}{W \cdot \left(\sum X^2\right) - \cdot \left(\sum X\right)^2} = \frac{< X^2 > \cdot < Y > - < X \cdot Y > \cdot < X >}{\left(< X^2 > - < X >^2\right)} , \tag{C.6}$$

o bien

$$b_0 = < Y > - a_0 \cdot < X > . \tag{C.7}$$

Donde usamos la notación:

$$\sum Y \equiv \sum_{i=1}^{N} w_i \cdot Y_i , \quad \sum X^n \equiv \sum_{i=1}^{N} w_i \cdot X_i^n , \quad \text{y} \quad \sum Y \cdot X \equiv \sum_{i=1}^{N} w_i \cdot Y_i \cdot X_i , \tag{C.8}$$

y así sucesivamente. También definimos los valores medios de Y y de X como:

$$\overline{Y} \equiv < Y > \equiv \frac{1}{W} \sum_{i=1}^{N} w_i \cdot Y_i , \quad \text{y} \quad \overline{X} \equiv < X > \equiv \frac{1}{W} \sum_{i=1}^{N} w_i \cdot X_i , \tag{C.9}$$

Por su parte, las desviaciones estándar y varianza vienen dadas por:

$$S_x^2 \equiv \frac{1}{W} \sum_{i=1}^{N} w_i \cdot \left(X_i - \overline{X}\right)^2 = \left(\frac{N-1}{N}\right) \cdot Var(X) = < X^2 > - < X >^2 , \tag{C.10}$$

y

$$S_Y^2 \equiv \frac{1}{W}\sum_{i=1}^{N} w_i \cdot \left(Y_i - \overline{Y}\right)^2 = \left(\frac{N-1}{N}\right) \cdot Var(Y) = < Y^2 > - < Y >^2 . \qquad \text{(C.11)}$$

Los coeficientes de correlación se definen en modo similar:

$$S_{YX} \equiv Cov(Y,X) = \frac{1}{W}\sum_{i=1}^{N} w_i \cdot \left(X_i - \overline{X}\right) \cdot \left(Y_i - \overline{Y}\right) = < YX > - < Y >< X > , \qquad \text{(C.12)}$$

y

$$R \equiv \frac{< YX > - < Y >< X >}{S_X \cdot S_Y} = \frac{Cov(Y,X)}{S_X \cdot S_Y} . \qquad \text{(C.13)}$$

El error típico de estimación de *Y* sobre *X* está relacionado con el valor de chi cuadrado normalizado, χ_N^2, por:

$$\chi_N^2 = \frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} = Error.típico(YX)^2 . \qquad \text{(C.14)}$$

También es útil definir el valor de chi cuadrado por grados de libertad, χ_v^2:

$$\chi_v^2 = \frac{N}{(N-2)} \cdot \frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} = \frac{N}{N-2} \cdot \chi_N^2 . \qquad \text{(C.15)}$$

La varianza total, $S_t \equiv S_Y$, da una medida de cómo los puntos Y_i se distribuyen alrededor del valor medio de *Y*. S_t se define como:

$$S_t^2 \equiv \frac{1}{W}\sum_{i=1}^{N} w_i \cdot \left(Y_i - \overline{Y}\right)^2 = S_Y^2 . \qquad \text{(C.16)}$$

La varianza inexplicada, S_{inex}, mide la calidad del modelo lineal propuesto, $Y(X_i)=aX_i+b$, para explicar los datos observados, (Xi,Y_i). Este nombre surge del hecho de que si el modelo lineal propuesto fuese adecuado, los valores $\varepsilon_i=(Y_i-Y(X_i))$ deberían tener una distribución estadística al azar. S_{inex} se define como:

$$S_{inex}^2 \equiv \frac{1}{W}\sum_{i=1}^{N} w_i \cdot \left(Y_i - Y(X_i)\right)^2 = \chi_N^2 . \qquad \text{(C.17)}$$

La variación explicada, S_{exp}, se define por:

$$S_{\exp}^2 \equiv \frac{1}{W}\sum_{i=1}^{N} w_i \cdot \left(Y(X_i) - \overline{Y}\right)^2 \qquad \text{(C.18)}$$

Sumando las Ecs. (C.16), (C.17) y (C.18) miembro a miembro, se puede demostrar que:[5,6]

$$S_t^2 = S_{\exp}^2 + S_{inex}^2 , \qquad \text{(C.19)}$$

de donde tenemos:

$$R^2 = \frac{S^2_{\exp l}}{S^2_t} = 1 - \frac{S^2_{inex}}{S^2_t} = 1 - \frac{\chi^2}{N \cdot S^2_Y} \leq 1 . \qquad \text{(C.20)}$$

Una propiedad importante de los estimadores a y b es que si los errores o residuos de las estimaciones, ε_i:

$$\varepsilon_i^2 = \left(Y_i - Y(X_i)\right)^2 \qquad \text{(C.21)}$$

tienen una distribución normal, entonces los valores a y b tendrán a su vez una distribución estadística y sus valores medios vendrán dados por $<a>=a_0$ y $<b>=b_0$ y sus desviaciones estándares, o *errores estándares*, denotadas por Δa_0 y Δb_0 respectivamente, son:[4]

$$\Delta a_0 = \frac{a_0}{\sqrt{N-2}} \cdot \sqrt{\left(\frac{1}{R^2} - 1\right)} \qquad \text{(C.22)}$$

y

$$\Delta b_0 = \Delta a_0 \cdot \sqrt{< X^2 >} . \qquad \text{(C.23)}$$

Si los errores ε_t tienen una distribución normal, la variable aleatoria t, definida por:

$$t = \frac{(a - < a >)}{\Delta a_0}, \qquad \text{(C.24)}$$

presentará una distribución t-Student,[6,7] con N-2 grados de libertad.

Para calcular la incerteza en la estimación de a_0 (Δa) a partir de una muestra de tamaño N, con un límite de confianza de $P\%$, se calcula a partir del valor t_p, que se obtiene de la distribución t-Student con (N-2) grados de libertad: [6,7]

$$Probabilidad\_t\text{-}Student\ (t<t_p) = P\%. \qquad \text{(C.25)}$$

Si se usa Excel® Microsoft, este valor de t_p se calcula usando la función DISTR.T.INV((1-P), N-2). La incerteza Δa se calcula como:

$$\Delta a(P\%) \equiv \Delta a_p = t_p \cdot \Delta a_0 = t_p \cdot a_0 \cdot \sqrt{\frac{(1/R^2 - 1)}{(N-2)}} . \qquad \text{(C.26)}$$

El error en b_0 viene dado por:

$$\Delta b(P\%) \equiv \Delta b_p = \Delta a(P\%) \cdot \sqrt{< X^2 >} . \qquad \text{(C.27)}$$

C.2 Significación estadística de parámetros de un ajuste

Un ensayo usual y necesario es evaluar o determinar si el valor de una pendiente u otro parámetro obtenido de un dado experimento es significativamente distinto de cero o

no. En definitiva, lo que deseamos evaluar es la hipótesis nula [#]H_0: $a_0=0$ frente a la hipótesis H_1: $a_0\neq 0$. En este caso analizamos si el intervalo definido por (a_0-Δa_0, a_0+Δa_0) contiene o no el valor cero. Si este intervalo contiene a cero, es claro que un valor de $a_0=0$ es consistente con el resultado y por lo tanto el resultado $a_0\neq 0$ no es estadísticamente significativo. El mismo criterio puede aplicarse para cualquier otro parámetro, por ejemplo b_0 —Ec. (C.23)—.

En el caso en ambas variables, X e Y, tengan errores, en general no hay un paradigma aceptado y las técnicas estadísticas para considerar estos casos son objeto de discusión entre los distintos autores y expertos en este tema. Aquí solo daremos algunas referencias para que el lector interesado las consulte en las Refs. [7], [8] y [9].

C.3 Caso de datos con error en las dos variables

Caso de error en ambas variables: En general las técnicas estadísticas para considerar estos casos son motivo discusión entre los distintos autores y expertos en este tema. Aquí proponemos un esquema aproximado, basado fundamentalmente en las Refs.[7], [8] y [9].

Si las mediciones (x_i,y_i) tienen errores, Δx_i y Δy_i respectivamente, y hay indicios de que la relación entre x e y es lineal, de la forma $y=a.x+b$, definimos los factores de peso para cada punto como:

$$W_i = 1/\sigma_i^2 , \tag{C.28}$$

donde:

$$\sigma_i^2 = a_0^2 \cdot \Delta x_i^2 + \Delta y_i^2 . \tag{C.29}$$

En general, si los factores de ponderación de la variable x e y son $w_{x,i}$ y $w_{y,i}$ respectivamente, entonces:

$$w_i = \frac{w_{x,i} \cdot w_{y,i}}{a_0^2 \cdot w_{y,i} + w_{x,i}} , \tag{C.30}$$

donde a_0 es la pendiente de la recta de regresión —Ec. (C.5)—. La dificultad es que para determinar a_0 debemos de resolver el problema de regresión. Para ello necesitamos los factores de peso w_i, que a su vez dependen de a_0. Para resolver este problema podemos proceder de modo iterativo. Usamos como ponderación inicial sólo los valores de $w_{yi}(=1/\Delta y_i^2)$. Con estos coeficientes, usando la Ec. (C.5) obtenemos el valor de a_0, y con este valor calculamos los pesos w_i; posteriormente, usando la Ec. (C.30) determinamos de nuevo los coeficientes w_i y a partir de la Ec. (C.5) los nuevos coeficientes a_0. Iterando hasta que los sucesivos valores de a_0 no cambien significativamente, se obtienen los parámetros de la regresión lineal buscada, o sea, la regresión lineal para el caso de datos con errores en las dos variables.

Estas ideas pueden extenderse al caso no lineal, en que la función $f(x;a,b,c,...)$, cuyos parámetros a, b, c,... se busca determinar, depende de un modo no lineal de x. En este caso la generalización de la Ec. (C.30) conduce al concepto de error efectivo:[8]

[#] Muchas veces formulamos una hipótesis con el único objeto de rechazarla. Por ejemplo, si deseamos decidir si una moneda está cargada o trucada, formulamos la hipótesis de que la moneda es buena. Estas hipótesis se denominan *hipótesis nulas*[5] y se designan con H_0. La máxima probabilidad con que deseamos rechazar una hipótesis cuando debió se aceptada (*Error tipo I*) se llama *nivel de significación* y se designa con α. Valores frecuentes de α son 0,005 (5%) ó 0,01 (1%).

$$\sigma_i^2 = \left(\frac{df}{dx}\right)^2 \cdot \Delta x_i^2 + \Delta y_i^2 \,. \qquad \text{(C.31)}$$

Referencias

[1] P. Bevington y D.K. Robinson, *Data reduction and error analysis for the physical sciences*, 2ª ed. (McGraw-Hill, New York, 1993).
[2] Stuardt L. Meyer, *Data analysis for scientists and engineers* (John Willey and Sons, Inc., New York, 1975).
[3] D.C. Baird, *Experimentación*, 2ª ed. (Prentice Hall Hispanoamericana S.A., México, 1991).
[4] J. Higbie, "Uncertainty in the linear regression slope", *Am. J. Phys*. **59**, 184 (1991).
[5] M. Spiegel, *Estadística*, 2ª ed. (McGraw-Hill, Bogotá, 1997).
[6] M. Spiegel, J. Schiller y R. Srinivasan, *Schaum's Outline of Probability and Statistics*, 2a ed. (McGraw-Hill, New York, 2000).
[7] J. Orear, "Least squares when both variables have uncertainties", *Am. J. Phys*. **50**, 912 (1982).
[8] D. Barker y L.M. Diana, "Simple method for fitting data when both variables have uncertainties", *Am. J. Phys*. **42**, 224 (1974).
[9] B. Cameron Reed, "Linear least-squares fits with errors in both coordinates II: Comments on parameter variances", *Am. J. Phys*. **60**, 1, (1992).